Published online 3 May 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W41–W45***
doi: 10.1093/nar/gkw330

SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species

Daniel Veltri[1,2], Martha Malapi Wight[1] and Jo Anne Crouch[1,*]

[1]Systematic Mycology and Microbiology Laboratory, U.S. Department of Agriculture (USDA), Agricultural Research Service (ARS), 10300 Baltimore Avenue, Building 10A, Beltsville, MD 20705, USA and [2]Oak Ridge Institute for Science and Education ARS Research Program, MC-100-44 P.O. Box 117, Oak Ridge, TN 37831, USA

Received February 19, 2016; Revised April 11, 2016; Accepted April 17, 2016

ABSTRACT

Defining syntenic relationships among orthologous gene clusters is a frequent undertaking of biologists studying organismal evolution through comparative genomic approaches. With the increasing availability of genome data made possible through next-generation sequencing technology, there is a growing need for user-friendly tools capable of assessing synteny. Here we present SimpleSynteny, a new web-based platform capable of directly interrogating collinearity of local genomic neighbors across multiple species in a targeted manner. SimpleSynteny provides a pipeline for evaluating the synteny of a preselected set of gene targets across multiple organismal genomes. An emphasis has been placed on ease-of-use, and users are only required to submit FASTA files for their genomes and genes of interest. SimpleSynteny then guides the user through an iterative process of exploring and customizing genomes individually before combining them into a final high-resolution figure. Because the process is iterative, it allows the user to customize the organization of multiple contigs and incorporate knowledge from additional sources, rather than forcing complete dependence on the computational predictions. Additional tools are provided to help the user identify which contigs in a genome assembly contain gene targets and to optimize analyses of circular genomes. SimpleSynteny is freely available at: http://www.SimpleSynteny.com.

INTRODUCTION

Understanding patterns of conserved synteny from the genomes of different organisms is a central undertaking in the field of molecular biology. Originally synteny was defined through cytogenetics, and referred to the presence of two or more loci located on a single chromosome (1). With the widespread application of next-generation sequencing and the ability to routinely assemble whole genome datasets, it is now also common to describe synteny interchangeably with collinearity, or the conservation of gene order and orientation. Qualitatively, synteny can take different forms, depending on the scale involved. Macrosynteny refers to collinearity of gene order at the whole-chromosome scale, microsynteny describes a small number of genes exhibiting collinearity across a given subchromosomal region and mesosynteny is characterized by the conservation of gene content within a chromosome in the absence of collinearity (2). Because the degree of synteny breaks down over time through various processes, including chromosomal rearrangements, gene losses and gains, and chromosomal duplications and losses, assessments of synteny allow biologists to address questions related to the evolutionary divergence of organisms and gene families. The focus of syntenic exploration may be limited to genes on a single chromosome, expanded to consider the organization of an entire genome and may be performed between multiple species depending on the question at hand (3).

Computational tools for assessing synteny are becoming increasingly important components of comparative genomic studies, particularly with the proliferation of draft genome assemblies containing large numbers of contigs. A few examples of programs for detecting novel syntenic regions include: Proteny (4), i-ADHoRe (5) or DRIMM-Synteny (6). A recent summary of additional predictive software tools is provided in (7) while a review on visualizing genomic comparisons is available in (8). Generally speaking, each program for estimating synteny offers a trade-off with respect to the number of genomes accommodated, method of scoring, scale of analysis (micro- versus macro-syntenic) and user interface (command-line versus graphical/web-based). When selecting a program for a particular project, it has been noted that some perform better when dealing with closely-related taxa, while others do better with more

*To whom correspondence should be addressed. Tel: +1 301 504 6922; Fax: +1 301 504 5062; Email: joanne.crouch@ars.usda.gov
Present address: Martha Malapi Wight, Plant Germplasm Quarantine Program, Animal and Plant Health Inspection Service, USDA, 9901 Powder Mill Road, Bldg. 580, Beltsville, MD 20705, USA.

divergent ones (4,7). Accordingly, results on the same data set can vary widely between programs (9).

Programs typically provide visual assessments of macrosynteny in the form of dot or circle plots, which often lack the fine scale level of detail necessary to easily observe individual genes. For example, the web-based CoGe Comparative Genomics Platform (http://genomevolution.org) provides the SynMap tool (10) to generate a dot plot, which allows the user to click on a region of a chromosome to zoom in. It is even possible to click on a chromosome segment and be taken to CoGE's Genome Evolution Analysis (GEvo) viewer to display individual gene information if an additional GFF file has been provided. However, when a user hovers their mouse over a region of the dot plot in SynMap, no gene names are displayed. A user looking for an individual gene is required to estimate its general location and repeatedly jump back and forth between SynMap and GEvo until it is found. Another web-based visualization tool with greater focus on displaying syntenic information for individual genes across one or more genomes is the Multi-Genome Synteny Viewer (mGSV) (11). The program expects the user to identify their genes of interest prior to use and requires them to provide the exact location of each gene for all genomes. To help with this task, the authors provide Perl scripts to export results from BLAST (12). In turn, this requires the user to be familiar with the command line and could entail writing a custom script if there is a need to convert results from an unsupported format. The program also does not currently allow the user to save a figure of their final analysis. Since gene names are only displayed when the mouse hovers over it, taking a screenshot is not an ideal solution.

Keeping the above issues in mind, we present here a new web-based tool called SimpleSynteny to provide informative visuals of microsynteny. In contrast to the higher-level view provided by dot and circle plots, our pipeline provides a more detailed perspective for researchers exploring a preselected set of gene targets. Unlike dedicated browsers that display synteny for a subset of curated genomes, such as the Yeast Gene Order Browser (13), we allow the user to upload a limited number of contigs from any genome of their choosing. An emphasis has been placed on accessibility so that the tool is readily usable to those without advanced computer or scripting skills. In the following section we detail how the server works and highlight some features and additional tools before showing two examples of SimpleSynteny analyses. First, we show a side-by-side comparison of SimpleSynteny and mGSV by recreating an analysis of a secondary metabolite cluster in two fungi. The second example details a more complex analysis, using mating genes found across eight different fungi.

MATERIALS AND METHODS

Workflow overview

The standard pipeline for SimpleSynteny consists of three primary steps: (1) data input, (2) contig editing and (3) customization of graphical output. Users are also provided with two optional tools that identify contigs of interest from FASTA files, and an advanced mode that allows for custom image manipulation.

Data input. The SimpleSynteny pipeline begins on the main 'Step 1' page where individual genome and gene target files in FASTA format are uploaded. At least one genome file and one gene file are required, but a single gene target file can be assigned to multiple genomes as discussed below. FASTA definition lines are used to label all contigs and genes, however, the user can manually edit the names of genomes during the upload process. Each genome file can contain up to ten contigs (or supercontigs, scaffolds, chromosomes, etc.). For cases where the user does not know which contig(s) in a complete genome assembly contains their genes of interest, the genome assembly can be preprocessed using the optional 'Contig Finder' tool (detailed below) to quickly identify and export only the sequences containing the target region(s) into a single merged FASTA file. Each gene target file can contain up to 60 nucleotide or protein sequences. When comparing multiple genomes, SimpleSynteny draws connections between genes with identical names. Accordingly, consistent gene names and spelling of gene definition lines is important if multiple gene target files are used. After completing file uploads, the user can adjust additional settings using the 'Advanced Settings for Gene Matches,' allowing for customization of basic BLAST parameters, a threshold to exclude target sequences which do not have a minimum percentage of positions contained within BLAST hits and an optimization setting for aligning circular genomes.

Discovery versus visualization. The SimpleSynteny visualization pipeline does not explicitly score or evaluate syntenic relationships between targets. Accordingly, users should be aware of the difference between assigning a single gene target file to multiple genomes, versus the use of individual gene target files that uniquely correspond to each genome. In the former case, when BLAST is searching using sequences from a different genome, novel discovery is taking place. Such evaluations can be a fast and convenient first step for researchers to visually explore syntenic relationships between genomes. However, confirmation of results using additional tools may be required, particularly, when comparing distantly related species. In contrast, when specific target files are supplied for all genomes, SimpleSynteny functions strictly as a visualization tool, as the orthologous relationship between targets have already been confirmed by the user.

Contig editing. In 'Step 2' of the SimpleSynteny pipeline, BLASTN or TBLASTN are used to align nucleotide or protein targets onto the contigs of the first genome, respectively. The target-mapping process typically takes less than 30 s and the hit with the best *E*-value for a given target on a particular contig is used to set the strand direction when drawing annotations. A warning will appear if a given target gene does not map to the genome, along with the reason that mapping was not completed, allowing the user the opportunity to go back to Step 1 and adjust threshold settings as appropriate. A preview figure will then load on screen, showing all contigs for the first genome in a horizontal layout, including gene locations and orientations. Genes are uniquely color-coded for ease of visualization. The user can adjust the position and orientation of each contig by mov-

ing it left or right, or by removing individual genes at their own discretion before repeating this process for any remaining genomes. Clicking the 'Show Other Genomes' button allows the user to see previously-edited genomes alongside the current selection to aid in decision making. When editing multiple genomes, the 'Try to Optimize Contig Order' button will attempt to automatically arrange contigs in the same order as the first edited genome. The editing of contig positions and orientation continues until all genomes are processed.

Graphical output. When all genomes have been edited, 'Step 3' of SimpleSynteny allows the user to adjust an array of image settings, save the final completed figure, and generate summary data from the analysis. The 'Basic Image Settings' section allows the user to select from several standard image formats (PNG, JPG, GIF, EPS, TIFF or PDF), adjust the width and height of the image, and customize image resolution. Images can be drawn at low resolution for rapid visualization, or the user can increase the resolution up to 1200 dots per inch for publication-quality graphics. Under 'Genome Adjustments,' the user may alter the size and placement of genome names. The program also provides an option to automatically attempt to declutter the syntenic diagram by reordering genomes to minimize either the Euclidean distance of lines connecting genes or the number of arrows indicating changes in gene direction. The section 'Drawing Style' provides options for converting a figure to gray scale, adjusting the manner in which gene labels are displayed, or toggling a gene shading option to highlight regions along the full-length sequence covered or excluded by significant BLAST hits. This later feature can be useful as a quick visual indicator of sequence homology, for example when mapping proteins from one taxa onto another. The user can repeatedly adjust any of the above settings and generate a new preview image before deciding to download a final image file. They can also bookmark the URL of the 'Step 3' page and revisit their project for up to 72 h before it is deleted from the server. Generating a preview image typically takes less than a minute but requires more time as image dimensions and resolution settings are increased. Final full-resolution figures are provided for download in a ZIP file. The archive also contains other useful documentation, including server settings, lists of any unmapped genes, results provided from BLAST and a set of human-readable 'contig mapping' (CMAP) files for each genome for use with Advanced Mode as described below.

Contig Finder

Contig Finder is an accessory tool included with SimpleSynteny to allow easy identification and extraction of contigs containing gene targets within a genome. The user first uploads a single genome in FASTA format, up to 250 MB in size. Larger genome files will need to be split into parts and processed separately. After the file is uploaded, a text area appears where the user can paste nucleotide or protein target sequences to search the genome using BLAST. If any hits are found, results are sorted in descending order to show contigs containing the most number of hits first. The user can then add up to 10 contigs to an export list to save the sequences in FASTA format.

Advanced mode

Advanced Mode allows the user to utilize CMAP files to directly interface with the SimpleSynteny figure generation engine. This allows the user to fully customize gene shading, or make additional edits to figures generated in Regular Mode using the CMAP files provided with user output as a starting template. Additional genes or knowledge obtained from other programs, for example those to search for tandem repeats, can also be incorporated. In brief, each CMAP file describes a single genome, with each line detailing an individual contig. Gene entries each delineate the name, direction and start/stop coordinates for both gene and shading boxes. Complete details on the CMAP format are provided in the site documentation. The Advanced Mode interface is designed to auto-correct the spacing of elements and to provide basic hints when invalid CMAP lines are submitted. Once $1-10$ valid CMAP files have been submitted through the Advanced Mode interface, the user can immediately advance to 'Step 3' (described above) to generate their figure.

Implementation details

The SimpleSynteny website utilizes JavaScript, jQuery and the Bootstrap framework (http://getbootstrap.com) to allow users to easily drag-and-drop files or use interactive buttons. The web interface is compatible with most modern browsers utilizing HTML5. Back-end scripting uses PHP to process files and Ruby and the BioRuby vr. 1.5 (14) package to interface with BLAST using user-provided parameters. Gene inputs are first scanned and determined as nucleic acid or amino acid sequences before being mapped to genomes using BLASTN or TBLASTN, respectively. Gene coordinates produced by BLAST are next processed into a CMAP file (mentioned above) for each genome, before being passed to a Ruby script using the RMagick (http://rmagick.rubyforge.org) interface for the ImageMagick library (http://www.imagemagick.org) to draw figures. To optimize the use of space, the program collapses and annotates large contig regions lacking any mapped genes as seen in Figures 1A and 2. SimpleSynteny also includes a Demo Mode to help new users quickly become acquainted with the user interface. The demo allows users to walk through and recreate an analysis of a fungal secondary metabolite gene cluster. First time users of SimpleSynteny are typically able to generate the Demo Mode figure in <10 min.

RESULTS

Demo mode example: recreating a syntenic analysis of a secondary metabolite protein cluster

O'Connell *et al*. recently highlighted the organization of a polyketide synthase secondary metabolite gene cluster (*Colletotrichum graminicola* Cluster 18) in two fungal plant pathogens, as shown in Figure 1B (15). Cluster 18 contains 15 genes, most of which are upregulated during host infection by the Arabidopsis pathogen *Colletotrichum higginsianum*, but not by the maize

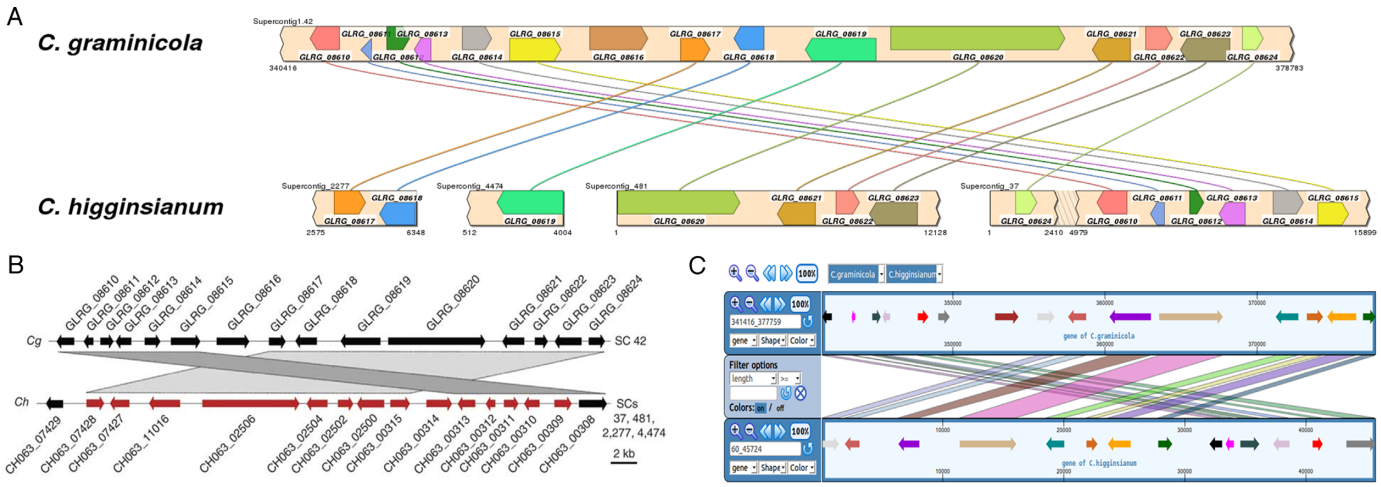

Figure 1. A comparison of syntenic analyses between a secondary metabolite cluster in the fungi *Colletotrichum graminicola* (*Cg*) and *Colletotrichum higginsianum* (*Ch*). (**A**) Syntenic diagram generated using SimpleSynteny's Demo Mode; which has the user map proteins from *Cg* onto both genomes. The protein sequence for *GLRG_08616* does not map onto *Ch* using default settings due to low sequence homology. Jagged edges and accompanying base pair numbers in *Ch* supercontig 37 denote the start and end of a contig region automatically collapsed by the program due to no genes being present to make the figure more compact. (**B**) The original figure from O'Connell *et al.* (15). Figure copyright of Nature Publishing Group and reused with permission. (**C**) Screenshot of the syntenic analysis performed using mGSV. Gene names are shown on the web-server version when highlighted by a mouse cursor.

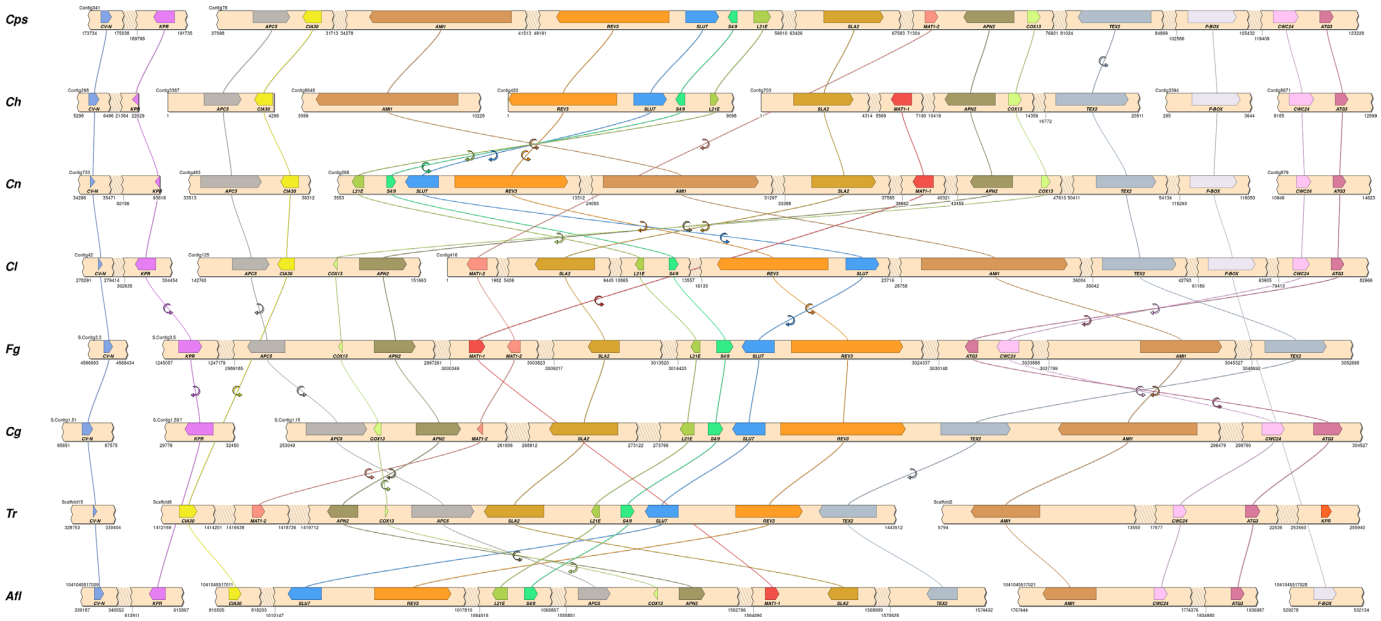

Figure 2. Syntenic analyses of the *MAT1* region for eight taxa within the Ascomycota. Homologous genes share a single color across species. Genome contig/scaffold segments are numbered with starting and ending nucleotide positions. Arrows represent changes in gene direction. Genes are drawn by using the protein representations from the genomes of *Calonectria pseudonaviculata* (*Cps*) and *Colletotrichum henricotiae* (*Ch*) and mapping them to genome assemblies using TBLASTN (*E*-value cutoff = 0.001). Shown are the following species: *Cps*, *Ch*, *Calonectria naviculata* (*Cn*), *Calonectria leucothoes* (*Cl*), *Fusarium graminearum* (*Fg*), *Colletotrichum graminicola* (*Cg*), *Trichoderma reesei* (*Tr*), and *Aspergillus flavus* (*Afl*). The question mark for *Cg* supercontig 1.59 has been manually appended to the figure to signify a possible substitution for another supercontig containing *KPR*.

pathogen *C. graminicola*. To recreate this syntenic analysis using SimpleSynteny, genomes and relevant proteins were obtained from the Broad Institute *Colletotrichum* Database (http://www.broadinstitute.org/annotation/genome/colletotrichum_group) for *C. graminicola* and *C. higginsianum*. Results generated using SimpleSynteny are shown in Figure 1A. For purpose of comparison, the same analysis was performed using mGSV, after genes were mapped using BLAST to help generate the required mGSV synteny and annotation files per the site's documentation (Figure 1C). Both programs reproduced the gene cluster organization, however, user time varied greatly, with SimpleSynteny taking ~10 min of user time versus ~45 min to prepare files using the mGSV pipeline. In addition, the quality of the output varied considerably, with SimpleSynteny yielding a customizable, publication quality graphic in either color or gray scale versus the mGSV screenshot.

A second example comparing the MAT1 locus between eight fungal taxa shows a conserved core group of genes in the region

In this example we show how SimpleSynteny is able to generate a more complicated syntenic comparison between the genomes of multiple, divergent organisms (Figure 2). Shown is a syntenic comparison of a genome region containing the fungal mating type gene *MAT1* and 16 surrounding genes, encompassing ~100-kb. These 17 genes were mapped to the genomes of eight filamentous fungi in the ascomycete sub-phylum Pezizomycotina, incorporating members of groups that last shared a common ancestor approximately 302 MYA (16). Depending on the organism, the 17 target genes were contained within 2 – 7 contigs. Detailed information describing this analysis and the datasets used to generate it are provided in supplemental materials.

DISCUSSION

We have presented here SimpleSynteny, what we hope to be a useful tool for biologists in a range of disciplines, including those without expertise working with command line software. The evaluation of structural changes among species is a fundamental step in many comparative genomics studies, and the visualization of intact, disrupted or duplicated gene regions is integral to many analyses. The SimpleSynteny pipeline is designed to provide a fast new method to quickly visualize syntenic gene regions mapped across one or more genomes. We envision the tool being used either independently, or in conjunction with other programs which specialize in broadly comparing entire genome assemblies. Researchers dealing with circular assemblies such as organelles or bacterial genomes may find SimpleSynteny particularly helpful, as our circular genome option automatically aligns genomes to start at the same gene without the need for editing files. In the future, we hope to incorporate additional features into SimpleSynteny, such as more ways to customize gene shading. We also hope to eventually add an option to highlight introns and exons through the server's Regular Mode. Additional details on how to use the SimpleSynteny tool are available in the website documentation and the server can be freely accessed without any login requirement at: http://www.SimpleSynteny.com.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Y. Rivera, C. Salgado Salazar, L. Beirn, J. Demers and the reviewers for their valuable feedback and suggestions to improve server functionality. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

FUNDING

2013-2015 United States Department of Agriculture (USDA)—Animal and Plant Health Inspection Service Farm Bill 10201 and 10007 Program funds (to J.A.C.); USDA—Agricultural Research Service (ARS) [project 8042-22000-279-00D]; USDA-ARS Floriculture and Nursery Research Initiative [project 0500-00059-001 to J.A.C.]. Inter-agency fellowship agreement between the United States Department of Energy (DOE) and the USDA through the Oak Ridge Institute for Science and Education ARS Research Participation Program Fellowship [DOE contract DE-AC05-06OR23100]; Class of 2013 USDA-ARS Headquarters Research Associate Award (to J.A.C.). Funding for open access charge: U.S. Department of Agriculture, Agricultural Research Service.
Conflict of interest statement. None declared.

REFERENCES

1. Renwick,J. (1971) The mapping of human chromosomes. *Annu. Rev. Genet.*, **5**, 81–120.
2. Hane,J.K., Rouxel,T., Howlett,B.J., Kema,G.H., Goodwin,S.B. and Oliver,R.P. (2011) A novel mode of chromosomal evolution peculiar to filamentous ascomycete fungi. *Genome Biol.*, **12**, R45.
3. Tang,H., Bowers,J.E., Wang,X., Ming,R., Alam,M. and Paterson,A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.
4. Gehrmann,T. and Reinders,M.J. (2015) Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level. *Bioinformatics*, **31**, 3437–3444.
5. Proost,S., Fostier,J., De Witte,D., Dhoedt,B., Demeester,P., Van de Peer,Y. and Vandepoele,K. (2012) i-ADHoRe 3.0 - fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.*, **40**, e11.
6. Pham,S.K. and Pevzner,P.A. (2010) DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, **26**, 2509–2516.
7. Tang,H., Bomhoff,M.D., Briones,E., Zhang,L., Schnable,J.C. and Lyons,E. (2015) SynFind: compiling syntenic regions across any set of genomes on demand. *Genome Biol. Evol.*, **7**, 3286–3298.
8. Nielsen,C.B., Cantor,M., Dubchak,I., Gordon,D. and Wang,T. (2010) Visualizing genomes: techniques and challenges. *Nat. Methods*, **7**, S5–S15.
9. Ghiurcuta,C.G. and Moret,B.M. (2014) Evaluating synteny for improved comparative studies. *Bioinformatics*, **30**, i9–i18.
10. Soderlund,C., Bomhoff,M. and Nelson,W.M. (2011) SynMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.*, **39**, e68.
11. Revanna,K.V., Munro,D., Gao,A., Chiu,C.-C., Pathak,A. and Dong,Q. (2012) A web-based multi-genome synteny viewer for customized data. *BMC Bioinformatics*, **13**, 190.
12. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Byrne,K.P. and Wolfe,K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, **15**, 1456–1461.
14. Goto,N., Prins,P., Nakao,M., Bonnal,R., Aerts,J. and Katayama,T. (2010) BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*, **26**, 2617–2619.
15. O'Connell,R.J., Thon,M.R., Hacquard,S., Amyotte,S.G., Kleemann,J., Torres,M.F., Damm,U., Buiate,E.A., Epstein,L., Alkan,N. *et al.* (2012) Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nat. Genet.*, **44**, 1060–1065.
16. Prieto,M. and Wedin,M. (2013) Dating the diversification of the major lineages of Ascomycota (Fungi). *PLoS One*, **8**, e65576.